Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS

Rachel Panckhurst, Catherine Détrie, Cédric Lopez, Claudine Moïse, Mathieu Roche, Bertrand Verine

Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., et Verine B. (2013). « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS ». Épistémè — revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques, 107-138.

Résumé :
Dans le cadre de cet article, on expose le déroulement du projet *sud4science* (www.sud4science.org). En premier lieu, on décrit la phase d'*acquisition* des données en provenance des SMS et du questionnaire, avant d'aborder les étapes successives d'*anonymisation*, de *transcodage* et d'*annotation* optionnelle. Ensuite, on présente les analyses (socio-)linguistiques des pratiques scripturales de l'*écriture SMS* (eSMS) qui ont débuté, ainsi que celles prévues à court et à moyen terme.

Mots-clefs : SMS en français, discours électronique médié, anonymisation, transcodage, annotation, analyses (socio-) linguistiques.

Abstract:
This article describes the *sud4science* project (www.sud4science.org). Firstly, the authors present the *acquisition* phase of both SMS data and questionnaire data. Secondly, they explain *anonymisation* techniques, *transcoding* and optional *annotation* phases. Finally, they propose preliminary (socio-) linguistic analyses of scriptural usage of *SMS writing*, and they also indicate those that are planned in the foreseeable future.

Key words: French SMS, mediated electronic discourse, anonymisation, transcoding, annotation, (socio-)linguistic analyses.

Introduction

Le SMS (*Service de messages succincts*) vient de fêter ses 20 ans, et le grand public s'en sert réellement depuis 14 ans[1]. Malgré l'arrivée sur le marché d'autres outils de communication, les textos n'ont pas encore été détrônés, bien que leur déclin futur soit régulièrement prédit. Tous les ans, des trillions de SMS sont encore échangés à travers le monde, et lorsque l'opportunité leur est offerte – selon le contexte, les pratiques individuelles et les habitudes générales en mutation constante – beaucoup d'utilisateurs préfèrent textoter à l'aide de leurs téléphones portables plutôt que de s'en servir pour parler. Analyser cette forme spécifique de *discours électronique médié* (Panckhurst, 2006) est primordial pour les chercheurs en sciences du langage, informatique, information-communication, psychologie, sociologie, etc. pour mieux comprendre de quelle(s) façon(s) la langue évolue, et pour observer d'éventuelles *mutations* en cours à tous les niveaux de la société. Cependant, les données émanant des SMS sont difficiles à analyser, car elles sont tout simplement problématiques à obtenir en quantité significative.

Depuis une dizaine d'années, des collectes de SMS ont néanmoins commencé, et ce dans différentes langues. En 2004, un groupe d'universitaires belges a démarré un projet

[1] Le premier SMS a été envoyé en décembre 1992. Les opérateurs téléphoniques n'ont incorporé l'utilisation pour le grand public que sept ans plus tard, en 1999 :
http://www.guardian.co.uk/technology/2012/dec/01/text-messaging-20-years;
http://www.lemonde.fr/technologies/article/2012/12/03/bon-anivrsr-chr-sms_1798797_651865.html.

international, intitulé *sms4science*, afin de recueillir, organiser (en une base de données mondiale), et analyser des SMS authentiques (www.sms4science.org, Fairon *et al.,* 2006, Cougnon, 2012). S'en sont suivies d'autres collectes de SMS : l'île de la Réunion (L.-A. Cougnon & G. Ledegen 2010), la Suisse (C. Dürscheid & E. Stark 2011), le Québec (Langlais *et al.* 2012), la région Rhône-Alpes en France (Antoniadis *et al* 2011)[2]. L'initiative la plus récente pour le français est le projet *sud4science LR* (www.sud4science.org). En trois mois, (15/9/11 au 15/12/11), plus de 90 000 SMS authentiques ont été recueillis auprès du grand public par un groupe de chercheurs dans la région Languedoc-Roussillon (Panckhurst & Moïse, 2012b, Accorsi *et al*., 2012, Patel *et al*., 2013).

Dans le cadre de cet article, nous allons expliquer le déroulement du projet *sud4science*, en commençant par l'*acquisition* des données (§ 1). Nous présenterons ensuite les phases d'*anonymisation*, de *transcodage* et d'*annotation* optionnelle (§ 2), avant de nous intéresser de plus près aux analyses (socio-)linguistiques qui ont débuté, et à celles qui sont prévues à court et à moyen terme (§ 3).

1. Acquisition des données

Les personnes désireuses de participer à la collecte de SMS se sont inscrites sur le site web du projet (http://www.sud4science.org). Un questionnaire sociolinguistique était proposé et 95 % des participants l'ont complété. Parmi les personnes qui ont rempli à la fois le formulaire de consentement (obligatoire pour faire don de ses SMS par la suite) et le questionnaire (obligatoire seulement pour ceux qui voulaient participer à un tirage au sort hebdomadaire), 424 personnes ont effectivement fait don de leurs SMS. 93 085 messages ont été recueillis pendant la période de 13 semaines, et chaque donateur a envoyé 214 messages en moyenne. Le corpus épuré[3] contient 88 683 SMS en provenance de 424 personnes, soit, en moyenne, 209 textos par donateur (*cf.* Panckhurst et Moïse, 2012b). Les messages contiennent en moyenne 55 caractères, sans espaces, 67 caractères, avec espaces ; chaque message est d'une longueur moyenne de 13,75 mots. 14 677 SMS sont dotés de ≤ 15 caractères, soit 16,55 % du corpus *épuré*. Il est intéressant de comparer ces chiffres avec l'étude belge de 2004, dans laquelle le pourcentage est légèrement plus élevé (17,3 %) ; au contraire, on aurait pu s'attendre à ce que le pourcentage de messages très courts soit supérieur dans notre corpus de 2011, puisque les forfaits mensuels actuels incluent souvent les SMS illimités et, de ce fait, les scripteurs n'hésitent plus à envoyer des messages très courts (*e.g.* « ok ») en réponse à des SMS reçus.

1.1. Résultats préliminaires du questionnaire

Les participants sont jeunes, 80 % ont moins de 30 ans. Le pic de participation se situe à 18 ans, avec 33 donateurs distincts. Les plus jeunes sont des collégiens, en 6e, et la personne la plus âgée a 66 ans (un homme, niveau secondaire, CAP-BEP, retraité). Une seule personne a indiqué avoir un niveau d'études du primaire (il s'agit d'un homme, employé/ouvrier de 43

[2] Par ordre chronologique, après le recueil belge initial, les pays et régions suivants ont récolté des SMS authentiques, toujours dans le cadre du projet international *sms4science* :
Île de la Réunion : http://www.lareunion4science.org/ (20 000 SMS, 2008) ;
Suisse : http://www.sms4science.ch/ (24 000 SMS, 2009-2010) ;
Québec : http://www.texto4science.ca/ (5 000 SMS, 2010) ;
Grenoble: http://www.alpes4science.org/ (22,000 SMS, 2010).

[3] Les messages supprimés du corpus *sud4science* incluaient : des doublons, des messages envoyés par les responsables de la collecte aux participants, des SMS automatiques envoyés par les opérateurs téléphoniques, des textos en provenance de l'étranger, des messages reçus en provenance de personnes n'ayant pas rempli le formulaire de consentement, des messages publicitaires.

ans). Parmi les personnes inscrites, ayant rempli le questionnaire, plus de 60 % sont des femmes, et le public élèves/étudiants constitue un peu plus de 50 % ; parmi ce public estudiantin un pourcentage important (63 %) est en licence. Le fait que le public soit élargi montre peut-être que l'engouement médiatique a eu un impact bien au-delà du milieu universitaire et de l'école secondaire. Les types de téléphones nous ont vraiment interpellés également : 73 % des donateurs ont un téléphone intelligent, alors que notre étude a été menée à partir d'une université de sciences humaines, intégrant une proportion importante de boursiers. Enfin, 80 % des inscrits ayant rempli le questionnaire utilisent les SMS depuis plus de 5 ans.

En réponse à la question Pourquoi écrivez-vous des SMS (au lieu de téléphoner, par exemple ?) », les indications suivantes sont fournies, respectivement, en sachant que l'on pouvait cocher plusieurs cases :

- « parce que c'est moins cher ou inclus dans le forfait » (302 personnes) ;
- « pour aller plus vite » (294 personnes) ;
- « pour ne pas déranger » (211 personnes) ;
- « parce que je n'aime pas téléphoner » (143 personnes) ;
- « autre raison » (36 personnes) : « parce que ça laisse le choix au destinataire de répondre ou d'attendre suivant la situation dans laquelle il se trouve. » ; « [ce sont des] messages qui ne sont pas assez importants pour téléphoner à la personne » ; « pour pouvoir les relire » ; « le plaisir d'écrire ».

À la question « Si vous rédigez en écriture SMS, pourquoi le faites-vous ? », voici les réponses obtenues :

- « parce que c'est plus rapide » (293 personnes) ;
- « parce que ça crée une complicité entre amis » (58 personnes) ;
- « parce que j'aime jouer avec la langue » (42 personnes) ;
- « autre » (4 personnes) : « j'écris souvent de long sms, dc ça raccourcis sans diminuer le contenu », « c'est amusant decrire certains mots d'une certaine maniere; en plus d'être plus rapide on peut presque déterminer qui ecrit tel sms avec des mots particuliers ecrits d'une manière particulière et c'est amusant ^^ [4]».

Notre questionnaire, qui doit être davantage analysé dans un avenir proche, montre que les SMS sont voués à un avenir encore prometteur, en incluant une dimension de *discrétion* ou de choix du *moment de communication* (« Ils sont très pratiques car ils me permettent de faire passer un message sans déranger la personne. On doit répondre tout de suite (plus ou moins) à un appel alors que nous pouvons répondre aux sms quand nous voulons. »), ou sans s'encombrer de *rituels conversationnels* (« L'envoi de sms permet de communiquer quasi instantanément en allant directement au but, contrairement à la conversation téléphonique qui induit un certain nombre de passages obligés avant d'entamer le sujet réel de conversation. »), ou encore en tant qu'*aide-mémoire* (« moi qui ai peu de mémoire, ça me permet de garder les infos (adresses, heure de rdv...) », ou en *se protégeant* derrière un écran (« écrire ce que l'on n'ose pas dire ») et enfin pour des publics ayant différentes spécificités (« je suis sourde, c'est plus simple! »). Toutes ces caractéristiques s'appuient sur les spécificités de la conversation par SMS, temps différé, moment d'échange choisi, etc.

[4] Les SMS sont retranscrits tels quels.

2. Anonymisation, transcodage, annotation

Lorsque le corpus *sud4science* aura été entièrement anonymisé, transcodé et (éventuellement) annoté, la fouille et l'analyse à grande échelle pourront débuter. Les tâches d'anonymisation et de transcodage sont cruciales pour la suite du projet et pour la mise à disposition du corpus.

2.1. Système d'anonymisation

Dans le but de masquer l'identité d'un individu, l'anonymisation se révèle une tâche indispensable, par exemple dans le domaine juridique (Plamondon *et al.,* 2004) ou médical (Grouin *et al.* 2009). Dans ces domaines, les systèmes reposent principalement sur la reconnaissance automatique des noms, des dates, des lieux et d'autres éléments qui peuvent conduire à l'identification des personnes. Généralement les méthodes de reconnaissance de ces types d'entités nommées s'appuient sur des règles spécifiques et l'utilisation de dictionnaires. De plus, des méthodes d'apprentissage supervisé peuvent être appliquées. Par exemple, (Szarvas *et al.,* 2007) ont entraîné plusieurs classifieurs afin de proposer une fonction de prédiction combinant les résultats pour une tâche d'anonymisation. Une limite essentielle de ces méthodes est liée à la nécessité de disposer d'une quantité importante de données étiquetées.

Nous considérons qu'un tel processus d'anonymisation ne peut être entièrement automatique. Suivant cette même hypothèse, les travaux de (Reffay *et al.*, 2012) se focalisent sur la création d'une interface par laquelle l'expert peut identifier les données personnelles et décider si elles nécessitent d'être anonymisées. Le logiciel d'anonymisation de SMS que nous proposons repose sur le même principe, en sachant que nous cherchons à faciliter le travail de l'annotateur par une procédure automatique. Dans de telles situations, les marqueurs identitaires des SMS doivent être anonymisés[5]. Par exemple, le SMS (*cf.* Figure 1) « G pas encore de rep de sab! […] », nécessite l'anonymisation du mot « *sab* ». Ici, la difficulté majeure réside dans le fait que ce marqueur est atypique et ne représente pas un prénom trivial, c'est-à-dire appartenant à un dictionnaire de prénoms et/ou de surnoms.

Le logiciel d'anonymisation *Seek&Hide,* développé par deux étudiants en informatique[6], s'appuie sur des méthodes de TALN (Traitement Automatique du Langage Naturel). Il propose une page web sécurisée accessible pour les annotateurs. Le but du logiciel est de faciliter l'expertise et de traiter une quantité importante de données. L'approche développée se décline en trois phases :

- ***Phase automatique*** : traitement automatique des données (mots) qui ne présentent *a priori* aucune ambiguïté quant à leur interprétation (à anonymiser ou non).
- ***Phase semi-automatique*** : traitement manuel de l'information nécessaire pour les SMS qui présentent des mots ambigus ou inconnus. Ceci s'effectue à travers un système qui met en relief les éléments nécessitant une expertise. Cette mise en valeur facilite significativement le travail de l'annotateur.
- ***Phase de validation*** : relecture et validation des SMS anonymisés automatiquement ou suppression d'une anonymisation incorrectement appliquée par l'outil lors de la phase automatique.

[5] Notre correspondant *Informatique et Libertés* (*cf.* CNIL, www.cnil.fr), du *Service des affaires juridiques et institutionnelles* de l'université Paul-Valéry Montpellier 3, Nicolas Hvoinsky, a rempli une déclaration exigeant que l'anonymisation de la totalité du corpus de SMS *sud4science* ainsi que des données émanant du questionnaire soit effectuée avant le 30/9/13.

[6] Pierre Accorsi et Namrata Patel (étudiants en Master Informatique) ont développé le logiciel d'anonymisation *Seek&Hide* pendant un stage de deux mois en 2012 (*cf.* Accorsi *et al*., 2012, Patel *et al*., 2013).

2.1.1 La phase automatique

De manière concrète, le traitement des données textuelles se décline en trois étapes successives que nous synthétisons ci-dessous :

A) Pré-traitement des données. La première étape consiste à segmenter le corpus en mots.
B) Identification des mots susceptibles ou non d'être anonymisés. Dans cette phase du processus automatique, chaque mot d'un texte peut demander d'être anonymisé (AA : à anonymiser) ou non (NPA : ne pas anonymiser). Pour cela, nous utilisons deux types de dictionnaires :

- un « Dictionnaire » qui contient des mots qui doivent être rendus anonymes. Le dictionnaire que nous utilisons est constitué d'une liste de prénoms.
- un « Anti-dictionnaire » qui contient des mots qui ne nécessitent pas d'anonymisation. Cet anti-dictionnaire est issu de la fusion de différentes ressources : lexique des formes fléchies du français (Lefff[7]), dictionnaire de certaines formes récurrentes utilisées dans l'écriture SMS (par exemple, les binettes, certaines abréviations, etc.), dictionnaire de noms de lieux.

C) Traitement des mots à anonymiser. Chaque mot est traité en vérifiant son appartenance aux différents dictionnaires et anti-dictionnaires. Quatre situations sont rencontrées comme l'illustre le tableau 1 ci-dessous :

Traitement du mot	**Dans le dictionnaire ?**	**Dans l'anti-dictionnaire?**	**Type**	**Traitement**
Rachel	Oui	non	Dictionnaire	Automatiquement anonymisé
crayon	Non	oui	Anti-dictionnaire	Ignoré (ne pas anonymiser, NPA)
Pierre	Oui	oui	Ambigu	Surligné (candidat pour la phase semi-automatique)
Namrata	Non	non	Inconnu	Surligné (candidat pour la phase semi-automatique)

Tableau 1 : Identification automatique des mots à traiter (*cf.* Accorsi *et al*, 2012)

Le mot *Rachel* doit être anonymisé car il apparaît dans le seul dictionnaire des prénoms ; le mot *crayon* est ignoré car il apparaît uniquement dans l'anti-dictionnaire (LEFFF) ; le mot *Pierre* est *ambigu* car il est présent dans les deux dictionnaires ; enfin, *Namrata* est *inconnu* car il est absent des deux dictionnaires. Dans ces deux derniers cas, le mot est surligné par le logiciel (*cf.* Figure 1) et sera à traiter dans la phase d'anonymisation semi-automatique.

Un exemple d'anonymisation de prénoms complétée figure dans le tableau 2 ci-dessous :

Coco est pas la ! Éva non plus ! Tanpis ! Lol J'irai aux journée du patrimoine ! Éva m'a dit que tu venais cette semaine peut etre ! Bisous ! !
<PRE_4> est pas la ! **<PRE_3>** non plus ! Tanpis ! Lol J'irai aux journée du patrimoine ! **<PRE_3>** m'a dit que tu venais cette semaine peut etre ! Bisous ! !

Tableau 2 : Du SMS « brut » au SMS anonymisé.
Les chiffres renvoient au nombre de caractères du prénom dans le SMS brut.

Cette section se concentre sur la phase d'anonymisation la plus complexe du processus, à savoir l'anonymisation des prénoms. Dans le cadre de nos travaux, d'autres types

[7] Lexique des Formes Fléchies du Français, LEFFF, http://alpage.inria.fr/~sagot/lefff.html

d'anonymisation[8] ont été réalisés qui s'appuient sur la mise en place d'expressions régulières pour identifier quelques éléments spécifiques (adresses de courriel, numéros de téléphone, adresses URL, etc.)

La partie la plus délicate du traitement automatique de ce type de corpus réside dans la mise en correspondance des dictionnaires compte tenu des spécificités lexicales des SMS. Ainsi, nous avons identifié les cas ci-dessous :

- Mots orthographiés, de manière non standard, par exemple : *surment* (à la place de *sûrement*)
- Mots écrits sans les accents, par exemple : *desole* (à la place de *désolé*)
- Mots avec des accents non standards, par exemple : *dèsolè* (à la place de *désolé*)
- Prénoms avec ou sans majuscules : *cédric* (à la place de *Cédric*)
- Répétition de lettres, par exemple : *nicooooolllaassss* (à la place de *Nicolas*)
- Formes abrégées, diminutifs : *Nico* (à la place de *Nicolas*)
- Onomatopées, par exemple : *mouhahaha*
- Élision sans apostrophe, par exemple : *jexplique* (à la place de *j'explique*)
- Agglutination, par exemple : *jtaime* (à la place de *je t'aime*)

Le tableau 3 résume les heuristiques (programmes informatiques) conçues pour traiter les cas cités ci-dessus. Notons que les différentes heuristiques développées donnent des résultats tout à fait satisfaisants car les situations décrites sont correctement reconnues de manière automatique dans plus de 96,9% des cas (Accorsi *et al*., 2012).

Nom	**Nom détaillé**	**Description**
WWoutA	Mots écrits sans signes diacritiques (*desole*)	Effectuer une désambiguïsation au moment de la recherche
WWithA	Mots écrits avec des signes diacritiques non standards (*dèsolè*)	Effectuer une désambiguïsation au moment de la recherche
OmiA	Élision (*jexplique*) et agglutination (*jtaime*)	Identifier et éliminer les préfixes tels que *jt*, *jl*, *j*, etc., puis effectuer la recherche
SRepet	Onomatopées (*mouhahaha*)	Détecter les répétitions de sous-chaînes telles que *ha*, *hé*
LRepet	Répétitions de caractères (*nicooooolllaassss*)	Supprimer les lettres identiques consécutives puis effectuer la recherche

Tableau 3 : Différents algorithmes pour le traitement de mots spécifiques aux SMS.

À partir des 88 683 SMS du corpus, *Seek&Hide* en a anonymisé 63 728 (soit 72 %) ; les 24 955 SMS restants (28 %) sont soumis à la phase semi-automatique suivante.

2.1.2. La phase semi-automatique

Seek&Hide propose une interface web sécurisée permettant aux annotateurs-experts linguistes[9] de mener à bien la phase suivante, qui permet de désambiguïser les SMS et de décider si l'anonymisation doit ou non être effectuée.

[8] Les étiquettes utilisées pour l'anonymisation sont les suivantes : Prénom (PRE), Nom (NOM), Surnom (SUR), Adresse (ADR), Lieu (LIE), Numéro de téléphone (TEL), Code (COD), URL (URL), Marque (MAR), Courriel (MEL), Autre.

[9] Deux étudiants de Master en Sciences du Langage, Camille Lagarde Belleville et Michel Otell, ont effectué cette phase, pendant trois mois (octobre-décembre, 2012).

Figure 1. Capture d'écran du logiciel *Seek&Hide*

La figure 1 montre les mots qui ont déjà été anonymisés pendant la phase automatique précédente (*Pauline* et *Lea*). Les autres mots surlignés requièrent une intervention humaine : *Juste* est NPA dans ce cas, mais le mot est surligné car il est potentiellement ambigu (tout comme *pierre*), entre un prénom et un mot figurant dans le dictionnaire LEFFF ; *elo* et *cece* sont AA, mais parce que ce sont des diminutifs, ils n'apparaissent pas dans le dictionnaire des prénoms ; *bebou* est anonymisé, car comme ce surnom n'est peut-être pas utilisé très souvent, il est potentiellement facile à reconnaître par des destinataires ou des tierces personnes ; *Penseea* montre les types de problèmes qui se posent et qui sont difficiles à traiter de manière automatisée, puisque l'espace est absente et correspond à : « Pense à » ; *Lec* est inconnu de tous les dictionnaires et dans ce cas est surligné ; cela est probablement une faute de saisie à la place de *Le*. Le dernier SMS montre d'autres mots modifiés, absents des dictionnaires : *petetre* à la place de *peut-être* (NPA), *surment* à la place de *sûrement*, (NPA), *pyis* à la place de *puis* (NPA), et *Juste*, encore une fois. L'extrait suivant (*cf.* figure 1), qui contient *ben*, présente en effet une ambiguïté (entre le prénom et l'interjection *bien*, abrégée dans ce contexte en *ben*) et doit être résolu grâce à un traitement humain : 'ya des gens beaucoup mieu placer que moi pour te comprendre et *ben* je peux essaye' ; dans la figure 1, *ben* est NPA.

Précisons que des approches d'apprentissage automatique issues du domaine de l'Intelligence Artificielle ont été développées et ont été combinées avec le système décrit dans cet article (Patel *et al*., 2013).

2.1.3. Phase de validation

La troisième phase[10] consiste en la lecture des SMS (72% du corpus) ayant été anonymisés de manière automatique par *Seek&Hide*, afin de vérifier si tous les textos ont bel et bien été correctement anonymisés. Trois cas de figure ont été repérés par les annotateurs :

1. *anonymisation automatique à enlever* :

[10] Deux étudiants de Master en Sciences du Langage, Frédéric André et Yosra Ghliss, ont effectué cette phase, pendant trois mois (février-avril, 2013).

grace a lui on comprend trop bien franchement ke kiffe la physique cette *anne* meme si cest bien dur

Dans cet exemple, *grace* et *anne* ont été anonymisés, mais ce n'est pas une erreur du logiciel. Si le scripteur avait ajouté l'accent circonflexe, Seek&Hide n'aurait pas procédé à l'anonymisation en prénom pour *grâce* ; l'autre occurrence est *anne* au lieu d'*année*, qui n'est donc pas un prénom dans ce cotexte.

2. *anonymisation manquante à insérer* :

Excuse pour c texto si tard c'était pour t dire q **mat** a u l permis bisous bisous

Mat est absent du dictionnaire de prénoms.

3. *balises d'anonymisation à remplacer :*

Une **clio** noir phase 2 vendue par une amie dune collègue de boulot.

Clio est ici un nom de voiture de la marque Renault, et non un prénom.

Les annotateurs humains peuvent donc retirer, ajouter, modifier les étiquettes précédemment insérées de manière automatique par le logiciel[11]. À ce stade, ils peuvent également décider de noter certains SMS comme devant être supprimés du corpus si ceux-ci contiennent des propos légalement inacceptables.

2.2. Transcodage et annotation

Une fois l'anonymisation terminée, les SMS sont prêts à être transcodés en français « standardisé » afin de permettre d'éventuels traitements ultérieurs en linguistique-informatique (incluant des analyseurs syntaxiques). L'idée est de restituer l'orthographe et la grammaire afin d'aider la compréhension, mais non pas d'« injecter » des éléments supplémentaires (*cf.* le tableau 4 ; par exemple, on n'ajoutera pas, dans cette occurrence, la particule de négation, *n'*). Les SMS bruts et transcodés seront tous disponibles après l'achèvement de cette phase. Le transcodage est utile pour le grand public, ou pour ceux qui veulent lire et comparer rapidement les SMS bruts et transcodés.

<PRE_4> est pas l**à** ! <PRE_3> non plus ! Tan**t** pis ! Lol**.** J'irai aux journée**s** du patrimoine ! <PRE_3> m'a dit que tu venais cette semaine peut-**ê**tre ! Bisous ! !

Tableau 4: SMS transcodé

Le corpus belge de 30 000 SMS a été manuellement transcodé (Fairon *et al*., 2007). Une fois cette opération effectuée, un système expert a été conçu pour aligner le corpus, caractère par caractère, et donc apprendre à partir des données, en comparant ainsi les SMS 'bruts' avec ceux qui avaient été transposés en français standardisé (Beaufort *et al*., 2010). Nous procéderons à une comparaison entre le traitement manuel et des méthodes semi-automatiques sur un échantillon de notre grand corpus.

Enfin, une phase d'annotation optionnelle prévoit l'utilisation de huit étiquettes : ABSence, BINettes, DIVers, GRAmmaire, LANgage, MODification, ORThographe, TYPographie. Quatre de ces étiquettes sont utilisées dans le tableau 5 ci-dessous (un double étiquetage peut être employé en cas d'ambiguïté) :

[11] Sur un échantillon de 20 000 SMS, seules 358 modifications ont dû être effectuées : 66 % (cas 1), 29 % (cas 2), 5 % (cas 3).

<PRE_4> **<ABS_n'>** est pas **<TYP_`>**là ! <PRE_3> non plus ! **<MOD_t_TYP_space>**Tant pis ! **<MOD_laughing out loud>**Lol **<TYP_.>**. J'irai aux **<GRA_MOD>**journées du patrimoine ! <PRE_3> m'a dit que tu venais cette semaine **<TYP_-^>**peut-être ! Bisous ! !

Tableau 5 : SMS annoté (étiquettes indiquées en gras)

3. Analyses (socio-) linguistiques

Bien que les phases d'anonymisation, de transcodage et d'annotation ne soient pas encore terminées, les analyses (socio-)linguistiques des SMS ainsi que des données émanant du questionnaire sociolinguistique ont déjà été initiées par les chercheurs.

Pourquoi étudier les SMS ?

Lors de la préparation de notre collecte, nous avons discuté à maintes reprises avec des personnes du grand public qui posaient systématiquement les questions suivantes aux linguistes de l'équipe : « Pourquoi voulez-vous étudier les SMS qui sont écrits dans un mauvais français ? » « Qu'est-ce que cela apporte à votre discipline, les sciences du langage ? » Nous répondions de la manière suivante : en tant que linguistes, nous observons, sans jugement, sans nous référer à une norme quelconque. Nous nous intéressons à l'étude du langage, des langues et des pratiques langagières, donc, lorsqu'il y a des mutations éventuelles, nous saisissons l'occasion pour étudier de nouveaux phénomènes. La langue est dynamique, en mouvance constante. Cette collecte a permis de recueillir en grand nombre des utilisations spontanées de la langue française. À partir de là, nous pouvons les comparer avec ce que nous connaissons d'autres types d'utilisations : par exemple, les écrits plus normés sur papier, ou encore les courriers électroniques, mais aussi les échanges oraux. Nous pourrons en tirer des conclusions sur la graphie, sur l'écriture, sur les choix de vocabulaire, sur la construction des phrases, sur la façon de s'adresser à l'autre, etc. Tous ces éléments varient selon les personnes, selon les âges et selon les situations. Par exemple, nous allons pouvoir étudier de quels mots se servent les gens pour entrer en contact (*salut*, *hello*, *coucou*) ou s'ils entrent immédiatement dans le vif du sujet ; ou encore, de quels mots se servent-ils pour solliciter leur destinataire. Vont-ils utiliser leur prénom, un mot doux, un surnom, etc. ? Puis, en étudiant les SMS, nous pouvons envisager des applications industrielles en TALN : des systèmes de transcodage SMS vers le français standardisé, pour des personnes qui ne comprendraient pas une écriture SMS très codée ; des logiciels de reconnaissance des SMS, ou de vocalisation, pour des personnes aveugles, ou pour des conducteurs, etc. Ces applications pourront aussi aider à traiter de grandes masses de données textuelles très présentes aujourd'hui dans les réseaux sociaux (tweets, Facebook).

Écriture SMS

L'écriture SMS[12] (désormais eSMS) est innovante, hautement créative. Pour mieux la saisir, Panckhurst (2009) a proposé une typologie (uniquement *néographique* et non *néologique*) des SMS pour le français, suite à d'autres chercheurs (Anis, 2004, Liénard, 2005, Fairon *et al.,* 2006, Véronis *et al.* 2006). Pour comprendre la complexité de l'eSMS, il fallait distinguer les cas de phénomènes *simples* (*substitution, réduction, remplacement, ajout, cf.* Tableau 6) et les cas de phénomènes *complexes*. Lorsque l'orthographe d'un lexème (*eau*) est totalement modifiée en une lettre (*o*), nous sommes face à une *substitution phonétisée entière* (phénomène *simple*, car une seule modification n'est effectuée). Mais, au sein des SMS, les phénomènes *complexes* abondent, par exemple *a2d1, 6T, 2manD* (*substitutions multiples*, dans ce cas), mais d'autres cas traversent les catégorisations, par exemple, des *substitutions*,

[12] Nous préférons *écriture SMS* à *écrit SMS* (Cougnon, 2012).

des *réductions* et des *suppressions* : *7éta* constitue une réduction graphique en agglutination et une suppression de fin de mot muette et une substitution phonétisée entière. Pour *chui, chais, yora, kestufé*, on a des agglutinations et du compactage et des écrasements. On peut rencontrer des cas d'ajouts/de variations, etc. : *moua*, *suuuupppppeer*.

substitution

phonétisée	entière : *o* (**eau**), *7* (**cet**)
	partielle : *ossi* (**au**ssi), *allé* (all**er**), bizes (bi**s**es)
	avec variation : *kikou* (c**ou**cou)
graphique	élision, typographie, majuscules : m en, est ce que
	icônes, symboles, rébus : *à* + (à **plus**), *de grandes* @ (**oreilles**)
	avec variation : *bisoux* (bisou**s**), *mwa* (m**oi**)

réduction

phonétisée	abrègements morpho-lexicaux : sigles/acronymes : *ASV, mdr, tvb, tlm, lol* troncations : *ordi* (ordinateur), *'lut, Net* (salut, Internet)
	variation : *ui* (oui), *i* (il)
graphique	fins/débuts de mots muet(te)s : *vou* (vou**s**), *peu* (peu**t**), *ôtel* (**h**ôtel) ; chute de e instables : *douch* (douch**e**)
	squelettes consonantiques & abréviations : *dc* (donc), *pr* (pour), *ds* (dans) ; consonnes doubles : *ele* (elle), *poura* (pourra) ; abréviations sémantisées : *t* (te/tu), *p* (peux/peut)
	agglutinations : jattends (j'attends)

suppression

graphique	typographie & ponctuation : […] *se genre de truc pr le site je pense ke ca devré allé vite je vou envéré […]*
	signes diacritiques : *ca* (ça), *voila* (voilà)

ajout

graphique	répétition (caractères, ponctuation) : *suuupppeer !!!!*
	représentations sémiologiques : :-)
	ajout de caractères : *oki* (ok), *les zamours* (les amours)
	onomatopées : *mouarf, arffff, bof*

Tableau 6 : Typologie de l'écriture SMS (Panckhurst, 2009)

3.1. Analyses des SMS

Quelques pratiques scripturales et leurs fréquences

Au sein du corpus *sud4science*, l'eSMS utilise souvent les *substitutions* et les *réductions*. Dans l'exemple suivant : « Slt j vé alé á la mèri 2m1 » : *Slt* = squelette consonantique, *j* = abréviation sémantisée, *vé/alé* = substitution phonétisée partielle, *á/mèri* = substitution graphique, *mèri* = suppression de fin de mot muette, *2m1* = substitutions phonétisées multiples (*cf.* Panckhurst, 2009 pour une description détaillée des termes utilisés). Des *augmentations/répétitions* et des *ajouts* de caractères, des *onomatopées* y figurent également (« *Grrr* ... On a réussit à changer nos billets ... On a pris le suivant ! On arrivera 1 heure après vous *!!! Pfffff* galère et fou rire *!!!* Et toi çà va *?!?* Dommage qu'on ai pas voyagé ensemble *!!!*. »). Le texto le plus court de notre corpus est « ok » ou « cc » (pour *coucou*) et le plus long contient un extrait de *La théorie de la relativité d'Einstein* et fait 4 658 caractères avec espaces. Des SMS n'incluant pas d'espaces existent :

1. […] tro.bisoutoucalinourienkepourtoipuissance 10<3
2. […] frontenormeetjouesdehamsterjovial

mais il est rare que ceux-ci constituent la totalité du message.

Les occurrences les plus élevées (en n-grammes) du corpus sont : *je* (36 153 1-gramme), « c est » (12 401, 2-grammes), « je t aime » (3 110, 3-grammes), suivi par « je sais pas » (1 414 et « c'est pas » (1 244).

Parmi les 30 000 occurrences de binettes utilisées, pour environ 50 binettes distinctes, dont les binettes textuelles les 5 plus fréquentes par ordre décroissant sont :) ^^ :p :d <3 seulement 1 % correspond à des binettes graphiques.

Plusieurs pistes et questionnements concernant les analyses (socio-)linguistiques sont actuellement à l'étude au sein de notre groupe de chercheurs :

Insultes-mots doux : Détrie et Verine (2012) constatent que les 'insultes' figurant au sein du corpus *sud4science* peuvent parfois être interprétées (de manière paradoxale) comme des 'mots doux'. L'intérêt est de montrer que le cotexte, parce qu'il entre en conflit dialogique avec l'insulte en question, permet d'interpréter cette dernière comme un mot doux, notamment grâce à des marqueurs de mise à distance de la dimension insultante de l'adresse et à leur surmarquage. Ils ont analysé 208 constructions d''insultes-mots doux' apparaissant au sein d'un échantillon important du corpus *sud4science* (22 500 SMS) :

- T'es impardonnable ! Mais j'taime quand meme *gros vilain*
- Bon anniversaire *sale HIPPIE*

Termes d'adresse : Détrie (2013) propose un classement et une analyse lexicale, discursive et énonciative des formes d'adresse statistiquement les plus fréquentes, et réfléchit aux nouvelles pratiques d'adressage au regard du genre SMS. Les *petit cœur, amour, poulette, mon amour doux, maman d'amour,* voire *mon petit poussin d'amour à moi en sucre de canne des îles* et autres mots caressants, auxquels il faut ajouter les surnoms et les prénoms déformés du type *ma Clarou* ainsi que les formes hypocoristiques du type *ma sœurette*, questionnent le genre en tant que ces apostrophes l'inscrivent encore très globalement dans la sphère privée, même si les usages évoluent actuellement, et qu'elles actualisent / spectacularisent un modèle énonciatif au sein duquel la dimension émotionnelle est prédominante.

Genre : Verine (2013) interroge les catégories de *genre* discursif *vs* textuel et la hiérarchisation de certains de leurs traits définitoires, à l'aune de la pratique récente du SMS et de ce qu'en écrivent ses usagers. Il analyse les énoncés métadiscursifs par lesquels les scripteurs de SMS commentent la forme, le contenu, les conditions de production ou d'interprétation des messages qu'ils écrivent ou reçoivent.

SMS isolés et SMS conversationnels : Une collecte de SMS conversationnels nous permettrait d'incorporer une étude incluant des interactions authentiques. Les contextes temporels et spatiaux pourraient être plus facilement explorés, en même temps que les rituels interactionnels, ainsi que l'analyse des *clôtures* et des *ouvertures* ; dans les SMS isolés du corpus *sud4science*, 75 % des items lexicaux étudiés correspondent à des clôtures (*à demain, à toute*, *bisous*, *bonne soirée*), contre seulement 25 % d'ouvertures (*Bonjour, bjr, hey, coucou, cc, ça va*). Cette pratique est-elle différente dans un corpus conversationnel, dans lequel les textos intermédiaires appartenant à un même 'fil' de discussion n'exigent pas nécessairement des ouvertures/clôtures nouvelles ?

Les types discursifs, les modes d'expression, les processus pragmatiques et discursifs constituent également des pistes pour des recherches futures : par exemple, comment exprime-t-on l'injonction dans les SMS ? Les modalisateurs sont-ils utilisés de manière aussi fréquente que lors de conversations en face-à-face ? Les traces d'émotion sont également importantes à analyser ; elles sont souvent indiquées par des représentations sémiologiques non-verbales (binettes), ou par des répétitions de caractères, de phonèmes ou de marques de ponctuation (*suuuuppeeerrr*, *ouiii, !!!!*) qui simulent l'intonation, c'est-à-dire de l'information paraverbale ; des interjections (*Hey*, *ben*), des mots du discours (*lol*) et des onomatopées (*grrr*) constituent d'autres exemples de ces traces.

D'autres pistes de recherche incluent la sémantique lexicale. En étudiant la créativité lexicale des textoteurs et les outils linguistiques au service de cette créativité, on pourrait classer les processus aboutissant à des innovations lexicales : dérivation propre (*miameur, dodoter, cinémater, facebooker, psychoter*), mots-valises (*chocobisous*), aphérèse (*un zou*), apocope (*Kiss sis :)*), emprunts plus ou moins francisés (*je cleane ; toi tu te fight pas beaucoup beaucoup hein*), modification de la construction verbale (par exemple construction transitive pour un verbe initialement intransitif : *j'ai zappe de prendre le paquet*), modification du genre pour les noms (*ma pote, ma chou*), innovation sémantique (*Tu vas lui dissoudre le porte monnaie pour des fringues au moins ?*), transformation d'une construction nécessairement personnelle en une construction non personnelle (*Sa fait quoi se soir?*), voire barbarisme délibéré (*j'ai mouru*), etc.

3.2. Liens entre SMS et données du questionnaire

Panckhurst et Moïse (2012a) évoquent le lien entre données et usages/pratiques. Se posent alors maintes questions : les utilisateurs de smartphones sont-ils toujours lexicalement créatifs, ou utilisent-ils systématiquement l'écriture intuitive, désormais intégrée dans le téléphone ? Les textos sont-ils plus longs maintenant qu'il y a dix ans ? Les forfaits mensuels incluant des SMS illimités contribuent-ils à des mutations quelconques ? L'âge des scripteurs est-il systématiquement un critère concernant le style d'eSMS ? Les outils de reconnaissance vocale (Siri, Iris, etc.) modifient-ils de manière importante les usages ? Les scripteurs multilingues font-ils souvent du code-switching quand ils rédigent des textos ? Un étudiant effectue actuellement une recherche sur les scripteurs qui se déclarent (dans le questionnaire) être bi ou trilingues, afin de comparer, d'une part, le lien entre leur(s) pratique(s) annoncée(s) et/ou leurs représentations de l'eSMS et leurs pratiques réelles, et, d'autre part, des différences éventuelles entre scripteurs monolingues et plurilingues.

Normes et écriture SMS

Moïse (2013) cherche à déterminer si les scripteurs reproduisent un discours normatif dans leurs textos, ou bien si l'eSMS créative les aide à renouveler leur conception de la norme. Le

fait de lier les données émanant du questionnaire et les SMS lui permettra de vérifier un éventuel écart entre la façon dont les scripteurs perçoivent leurs propres pratiques (représentations discursives dans le questionnaire) et leurs pratiques scripturales réelles (usage métalinguistique au sein des textos).

Conclusion

La phase d'*anonymisation* se termine avant le 30 septembre 2013 et le *transcodage* débute courant 2013. Une fois ces deux étapes primordiales terminées (voire l'étape d'*annotation*), le corpus sera organisé en une base de données, dont la diffusion est prévue auprès de chercheurs, d'étudiants et du grand public, en 2014.

Les analyses déjà initiées par les chercheurs en sciences du langage, en linguistique-informatique, et en informatique, à la fois à partir des données en provenance des SMS et de celles apparaissant dans les réponses au questionnaire, avec le croisement nécessaire entre les deux, pourront être menées sur l'ensemble du corpus.

Les pistes qui seront privilégiées en sciences du langage incluent : la sémantique lexicale, ou, plus précisément, la néologie lexicale et la néographie ainsi que les outils linguistiques au service de cette créativité (dérivation propre, mots-valises, aphérèse, apocope, emprunts, modification de la construction verbale, modification du genre pour les noms, innovation sémantique, transformation d'une construction nécessairement personnelle en une construction non personnelle, voire barbarisme délibéré, etc.) ; les termes d'adresse ainsi que les ouvertures et les clôtures ; les genres ; les types discursifs, les modes d'expression, les processus pragmatiques et discursifs.

La mise à disposition du corpus entre chercheurs de différentes disciplines permettra également un regard pluri-disciplinaire, qui est primordial pour bien appréhender toutes les facettes de ces données authentiques, photographie de discours médiés quotidiens de notre époque.

Remerciements : *Nous remercions la MSH-M (Maison des Sciences de l'Homme de Montpellier) et la DGLFLF (Délégation générale à la langue française et aux langues de France) qui soutiennent ce travail. Nous remercions vivement nos étudiants stagiaires : Anthony Stifani, étudiant en Master Information et Communication à l'Université Paul-Valéry Montpellier 3, qui a manuellement analysé une partie des SMS, permettant ainsi d'évaluer le système d'anonymisation ; Pierre Accorsi et Namrata Patel (étudiants en Master d'Informatique à l'Université de Montpellier 2), qui ont développé le système informatisé Seek&Hide, permettant d'anonymiser le corpus ; Michel Otell, Camille Lagarde-Belleville, Frédéric André et Yosra Ghliss (étudiants en Master de Sciences du Langage à l'Université Paul-Valéry Montpellier 3) qui ont procédé à l'anonymisation manuelle en ligne à l'aide de Seek&Hide et à la vérification de l'anonymisation automatique du corpus.*

Adresses

Rachel Panckhurst, Catherine Détrie, Bertrand Verine
Praxiling UMR 5267 CNRS & Université Paul-Valéry Montpellier 3
Route de Mende
F-34199 Montpellier cedex 5
rachel.panckhurst@univ-montp3.fr
catherine.detrie@univ-montp3.fr

bertrand.verine@univ-montp3.fr

Cédric Lopez
Objet Direct - VISEO
4, avenue Doyen Louis Weil
F-38000 Grenoble
clopez@objetdirect.com

Claudine Moise
Lidilem, Université Stendhal Grenoble 3
F- BP : 25 - 38040 Grenoble cedex 9
claudine.moise@u-grenoble3.fr

Mathieu Roche
LIRMM UMR 5506 CNRS & Université
Montpellier 2
161, rue Ada
F-34095 Montpellier Cedex 5
mroche@lirmm.fr

Références

Accorsi P., Patel N., Lopez C., Panckhurst R., Roche M. (2012), "Seek&Hide : Anonymising a French SMS corpus using natural language processing techniques", Lingvisticæ Investigationes, Special Issue : "SMS Communication : A Linguistic Approach", John Benjamins, 35:2, 163-180.

Anis, Jacques; Michel de Fornel; Béatrice Fraenkel. 2004 (organisers). « La communication électronique : Approches linguistiques et anthropologiques », Colloque international, EHESS, Paris, 5-6 February 2004.

Antoniadis G., Chabert G., Zampa V. (2011), « Alpes4science : Constitution d'un corpus de SMS réels en France métropolitaine », talk, 79th Acfas colloquium, Sherbrooke, 9-10 May 2011.

Richard Beaufort, Sophie Roekhaut, Louise-Amélie Cougnon, Cédrick Fairon (2010). "A hybrid rule/model-based finite-state framework for normalizing SMS messages". In: Hajič, Jan et al. (éds.): *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, Uppsala, Sweden, 11-16 July 2010. © 2010 Association for Computational Linguistics, pp. 770–779.

Cougnon L.-A., Ledegen G. (2010) « "c'est écrire comme je parle". Une étude comparatiste de variétés de français dans l'"écrit sms" ». *Les voix des Français. Modern French Identities*, 2(94):39–57.

Cougnon L.-A. (2012), « L'écrit sms. Variations lexicale et syntaxique en francophonie », PhD, Université Catholique de Louvain, September 2012.

Détrie C., Verine B. (2012), « Quand l'insulte se fait mot doux : la violence verbale dans les SMS », Colloquium *Dimensions du dialogisme 3 : Du malentendu à la violence verbale*, Helsinki, Finland, August 15-17, 2012.

Détrie C., (2013), « *Gentlemanminette d amour*, *ma chou*, *colocounette* et autres formes nominales d'adresse dans les SMS : de quelques spécificités liées au genre », Talk, Conference, « Interpréter selon les genres », April 18-20, 2013, Université Cadi Ayyad, Marrakesh, Morocco.

Dürscheid C., Stark E. (2011), "sms4science: An international corpus-based texting project and the specific challenges for multilingual Switzerland." In: C. Thurlow/K. Mroczek (eds.): *Digital Discourse. Language in the New Media.* Oxford: Oxford University Press, 299–320.

Fairon C., Klein J.-R., Paumier S., (2006), *SMS pour la science. Corpus de 30.000 SMS et logiciel de consultation*, Presses universitaires de Louvain, Louvain-la-Neuve, Manuel+CD-Rom, http://www.smspourlascience.be/

Fairon C., Klein J.-R., Paumier S., (2007), « Un corpus transcript de 30 000 SMS français », in Gerbault (éd.), *La langue du cyberspace : de la diversité aux normes*, L'Harmattan, p. 173-182.

Grouin C., Rosier A., Dameron O., Zweigenbaum P. (2009) « Une procédure d'anonymisation à deux niveaux pour créer un corpus de comptes rendus hospitaliers. » Risques, tech. de l'info. p. 23–34.

Langlais P. Drouin P., Paulus A., Rompré Brodeur E., Cottin F., (2012) "Texto4Science: a Quebec French Database of Annotated Short Text Messages" Proceedings, *LREC,* May, p.1047-1054.

Liénard F. (2005), « Langage texto et langage contrôlé. Descriptions et problèmes », *Linguisticae Investigationes,* Volume 28 n°1. Paris.

Moïse C. (2013), « "Lol non tkt on ta pas oublié" Rapports à la norme et valeurs de la « faute » dans l'écriture Sms (projet et corpus *Sud4science*). Réflexions sociolinguistiques.», « Si j'aurais su, j'aurais pas venu ! Linguistique des formes exclues : description, genre, épistémologie », Colloquium, Université Libre de Bruxelles, June 20-22.

Panckhurst R. (2006), « Le discours électronique médié : bilan et perspectives », in A. Piolat (Éd.). *Lire, écrire, communiquer et apprendre avec Internet*. Marseille : Éditions Solal, p. 345-366.

Panckhurst R., (2009), « Short Message Service (SMS) : typologie et problématiques futures. », in Arnavielle T. (coord.), *Polyphonies*, pour Michelle Lanvin, Université Paul-Valéry Montpellier 3, p. 33-52. [http://hal.archives-ouvertes.fr/hal-00443014].

Panckhurst R., Moïse C., (2012a), « sud4science Languedoc Roussillon, collecte de SMS isolés et conversationnels. Démarche et méthode scientifiques », communication, colloque VALS-ASLA, Lausanne, 1-3 février, 2012.

Panckhurst R., Moïse C., (2012b), "French text messages. From SMS data collection to preliminary analysis", Lingvisticæ Investigationes, Special Issue: "SMS Communication : A Linguistic Approach", John Benjamins, 35:2, 290-317.

Patel N., Accorsi P., Inkpen D., Lopez C., Roche M. (2013) "Approaches of anonymisation of an SMS corpus", Proceedings of CICLING (Conference on Intelligent Text Processing and Computational Linguistics), LNCS, Springer Verlag, March 24–30, 2013, University of the Aegean, Samos, Greece.

Plamondon L., Lapalme G., Pelletier F. (2004) « Anonymisation de décisions de justice », in Actes de *TALN'04*.

Reffay C., F. Blondel, E. Giguet *et al.* (2012) « Stratégies pour l'anonymisation systématique d'un corpus d'interactions plurilingues », in Actes du colloque *IC2012*, p. 1-21.

Szarvas, G., Farkas, R., Busa-Fekete, R. (2007) "State-of-the-art anonymization of medical records using an iterative machine learning framework", JAMIA 14(5) p.574-580

Verine B. (2013), « 'J VIENS DE ME FAIRE VIRER PAR SMS :(': Le sms est-il interprété comme un genre discursif ou textuel par ses usagers ? » Talk, Conference, « Interpréter selon les genres », April 18-20, 2013, Université Cadi Ayyad, Marrakech, Morocco.

Véronis J., Guimier de Neef E. (2006), « Le traitement des nouvelles formes de communication écrite », in Sabah, G. (Éd.), *Compréhension automatique des langues et interaction* (pp. 227-248). Paris : Hermès Science.

Référence à citer : Panckhurst R., Détrie C., Lopez C., Moïse C., Roche M., et Verine B. (2013). « Sud4science, de l'acquisition d'un grand corpus de SMS en français à l'analyse de l'écriture SMS ». Épistémè — revue internationale de sciences sociales appliquées, 9 : Des usages numériques aux pratiques scripturales électroniques, 107-138.